D1270–D1277 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 7 November 2011*
doi:10.1093/nar/gkr880

EcoliWiki: a wiki-based community resource for *Escherichia coli*

Brenley K. McIntosh[1], Daniel P. Renfro[1], Gwendowlyn S. Knapp[1], Chanchala R. Lairikyengbam[2], Nathan M. Liles[1], Lili Niu[1], Amanda M. Supak[1], Anand Venkatraman[1], Adrienne E. Zweifel[1], Deborah A. Siegele[3] and James C. Hu[1,*]

[1]Department of Biochemistry and Biophysics, Texas Agrilife Research, [2]Department of Computer Science and Engineering and [3]Department of Biology, Texas A&M University College Station, TX 77843, USA

Received August 14, 2011; Revised September 28, 2011; Accepted September 29, 2011

ABSTRACT

EcoliWiki is the community annotation component of the PortEco (http://porteco.org; formerly EcoliHub) project, an online data resource that integrates information on laboratory strains of *Escherichia coli*, its phages, plasmids and mobile genetic elements. As one of the early adopters of the wiki approach to model organism databases, EcoliWiki was designed to not only facilitate community-driven sharing of biological knowledge about *E. coli* as a model organism, but also to be interoperable with other data resources. EcoliWiki content currently covers genes from five laboratory *E. coli* strains, 21 bacteriophage genomes, F plasmid and eight transposons. EcoliWiki integrates the Mediawiki wiki platform with other open-source software tools and in-house software development to extend how wikis can be used for model organism databases. EcoliWiki can be accessed online at http://ecoliwiki.net.

INTRODUCTION

Laboratory *Escherichia coli* strains form the basis of much of our fundamental understanding of the molecular and genetic basis of life. As a central model system, *E. coli* has been either the primary focus or a key component for many bioinformatics data resources that cover overlapping but distinct aspects of *E. coli* biology. Many existing resources provide encyclopedic information about genes, gene products, transcripts and regulons of *E. coli* K-12 (1–6). Nevertheless, these resources only cover a fraction of the *E. coli* knowledge base wanted by biologists working with *E. coli*, which includes not only those interested in the biology of *E. coli per se*, but also many more using *E. coli* as a platform for a wide range of basic research and biotechnology.

We designed EcoliWiki (http://ecoliwiki.net) to facilitate community-driven sharing of biological knowledge of the model organism *E. coli*. First implemented in 2007, EcoliWiki is one of the early adopters of the wiki approach to model organism databases. Although it has been mentioned and cited in several reports and descriptions of other wiki-based projects (7–10), this report is the first comprehensive description of EcoliWiki.

One of our primary objectives for EcoliWiki is to capture community-contributed information about laboratory strains, phages, plasmids and so forth. Data about *E. coli* are being continually generated through bioinformatics, proteomic, genomic, biochemical and genetic research. This massive onslaught of data can be difficult to manage and integrate, though its importance to other scientists cannot be understated. It is crucial that the information being produced is interwoven with existing knowledge in an easily accessible and searchable resource. Linking all of this information to the relevant gene as well as to publications is vital for identifying or predicting the function of the gene product not only in *E. coli*, but also for orthologous gene products. EcoliWiki is designed with this linking and interoperability in mind.

*To whom correspondence should be addressed. Tel: +1 979 862 4054; Fax: +1 979 845 9274; Email: jimhu@tamu.edu
Present addresses:
Gwendowlyn S. Knapp, Wadsworth Center, NY State Department of Health, Albany, NY 12208, USA.
Lili Niu, Agilent Technologies, Santa Clara, CA 95051, USA.
Anand Venkatraman, Monsanto Company, 800 North Lindbergh Boulevard, St Louis, MO 63167, USA.

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

ECOLIWIKI CONTENT

EcoliWiki houses information in >64 000 content pages. As in other MediaWiki-based wikis, EcoliWiki content pages are paired with talk pages where users can discuss or debate content. EcoliWiki also extensively uses MediaWiki's system of placing pages in Categories, and the ability to nest Categories and Subcategories. Category terms increase the usability of the wiki for both users and automated processes.

Gene-centric pages

EcoliWiki contains groups of pages for genes from the *E. coli* lab strains MG1655 (11,12), W3110 (11), DH10B (13), BW2952 (14) and REL606 (15). Where a canonical gene name is available from EcoGene (5), it is used as the basis for the page names. Identifiers and non-canonical gene names are treated as synonyms that redirect to the gene, or to a disambiguation page when the same synonym is used for more than one gene. Orthologs from different strains are grouped together so that their annotations are made in a shared location. We used BLAST (16) to identify orthologs based on homology within the gene and in its flanking sequences. Note, however, that this method relies on orthologs being annotated as features in the GenBank or RefSeq files we use as sources. Orthologs that are present in a genome, but not annotated, will be missed. Improved projection of annotations between genomes is needed and is planned for the future. Some genes from the *E. coli* B strain REL606 share gene names with non-homologous genes in MG1655 (15). For example, two blocks of genes involved in LPS synthesis in MG1655 and REL606 are in similar locations. Several of these have common *waa* and *wbb* names although they do not have significant sequence similarity. To distinguish these from the K-12 genes, we give these genes from *E. coli* B their own sets of gene pages with page titles prefixed with E_coli_B. Other genes from REL606 that are not present in the K-12 strains do not have canonical gene names based on the Demerec system (17). In these cases, we base the gene names on the ECB locus tags (15).

In addition to *E. coli* genomes, we have gene-centric pages for the F plasmid, 21 different bacteriophage genomes and eight transposons (Table 1). These genes were imported from GenBank and RefSeq records, where available. The genome of bacteriophage phi80 and some of the transposon sequences were extracted from other sequence records. These pages are named with prefixes indicating the source genome, e.g. Phage_lambda_cI:Quickview.

Every gene in EcoliWiki is associated with six wiki pages linked by formatting that mimics tabs: *Quickview, Gene, Gene Product(s), Expression, Evolution* and *On One Page*. Each of the first five contains pertinent information about that aspect of the gene, and they all contain editable tables and notes. Each of these pages has a references section and a list of categories the page belongs to. EcoliWiki pages include links to other databases including EcoCyc (3), EcoGene (5), EchoBase (4), ASAP (2), EcoliGenExpDB (http://genexpdb.ou.edu/main/), RefSeq (18), UniProt (19), Pfam (20), Brenda (21), SwissModel (22), and ModBase (23). References are automatically generated using a modified version of the Cite extension (see below). The *On One Page* view catenates content from the other five for users who prefer scrolling to clicking between tabs. Figure 1 shows some of the elements of one of the gene-centric pages.

The *Quickview* page provides users with a brief summary of data in the form of a table that dynamically updates from tables on the other associated pages for the gene. The *Quickview* table also contains links to the DNA and protein sequences for the gene, literature searches and other database searches. Below the table, a *Notes* area is provided for general information. The *Gene*, *Gene Product(s)*, *Expression* and *Evolution* pages consist of multiple sections containing tables and *Notes* fields to allow capture of much more detailed information.

Information on the *Gene* page includes gene synonyms, mutant alleles and their availability, genetic interactions, and a set of images from GBrowse (25) showing the genomic context of the gene in each strain where it is found.

Table 1. Sources of genes listed on EcoliWiki

NCBI accession	Name	Notes
NC_000913	*E. coli* K-12 MG1655	
AC_000091	*E. coli* K-12 W3110	
NC_010473	*E. coli* K-12 DH10B	
NC_012759	*E. coli* K-12 BW2952	
NC_012967	*E. coli* B REL606	
NC_002483	Plasmid F	
NC_003298	Bacteriophage T3	
NC_000866	Bacteriophage T4	
NC_005859	Bacteriophage T5	
NC_001604	Bacteriophage T7	Extracted from a file (reannotated version of NC_001604) generously provided by Dr. Drew Endy
NC_001416	Bacteriophage lambda	
NC_002166	Bacteriophage HK022	
NC_002167	Bacteriophage HK97	
NC_001901	Bacteriophage N15	
NC_008720	Bacteriophage N4	
NC_005856	Bacteriophage P1	
NC_001895	Bacteriophage P2	
NC_001609	Bacteriophage P4	
	Bacteriophage phi80	Extracted from NC_010473
NC_001421	Bacteriophage PRD1	
NC_000929	Bacteriophage Mu	
NC_001422	Bacteriophage phiX174	
NC_003287	Bacteriophage M13	
J02448	Bacteriophage f1	
NC_001417	Bacteriophage MS2	
NC_001890	Bacteriophage Qbeta	
V00613	Transposon Tn3	
U00004	Transposon Tn5	
NC_002525	Transposon Tn7	Extracted from NC_002525
V00622	Transposon Tn9	
AF162223	Transposon Tn10	
V00359	Transposon Tn903	
V00612	Transposon Tn1681	
D16449	Transposon gamma-delta	

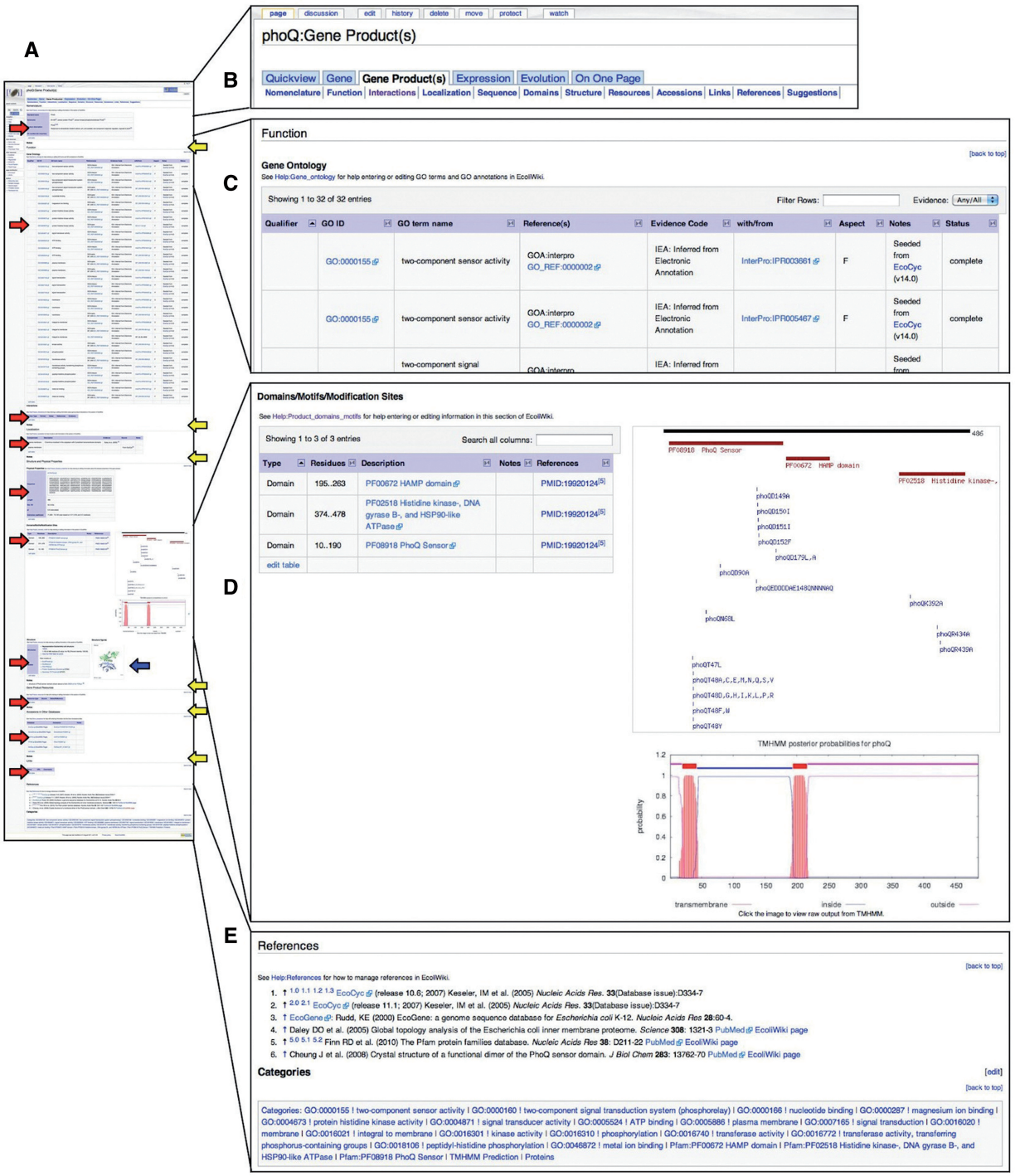

Figure 1. Example Gene-centric page for PhoQ product(s). (**A**) Overall page, showing the structure of coliWiki pages with tables (red arrows), figures, jmol viewer (blue arrow) and text notes (yellow arrows). (**B–E**) Show expanded views of page areas. (B) Simulated tabs for navigation between pages related to PhoQ, and links to page sections. (C) User-editable table for GO annotations for PhoQ. This table is periodically repopulated with annotations from EcoCyc. (D) Domains and motifs table and associated figures. The motifs diagram is generated from the content of the domains table to its left, and from the alleles and phenotypes table on the PhoQ:Gene page. The TMHMM (24) diagram is automatically generated for all genes. (E) References and categories. References are automatically generated from a software extension that recognizes PMIDs in the tables and embedded via markup in the text notes sections.

The *Gene Product(s)* page includes information about the protein or RNA. A major focus of EcoliWiki is the table for functional annotation using the Gene Ontology (GO) (26). This table allows users to add and correct GO annotations that we share with EcoCyc (8) and regularly deposit with the GO consortium (Figure 2). The *Gene Product(s)* page also includes sections for physical interactions, domains and motifs, structures, physical

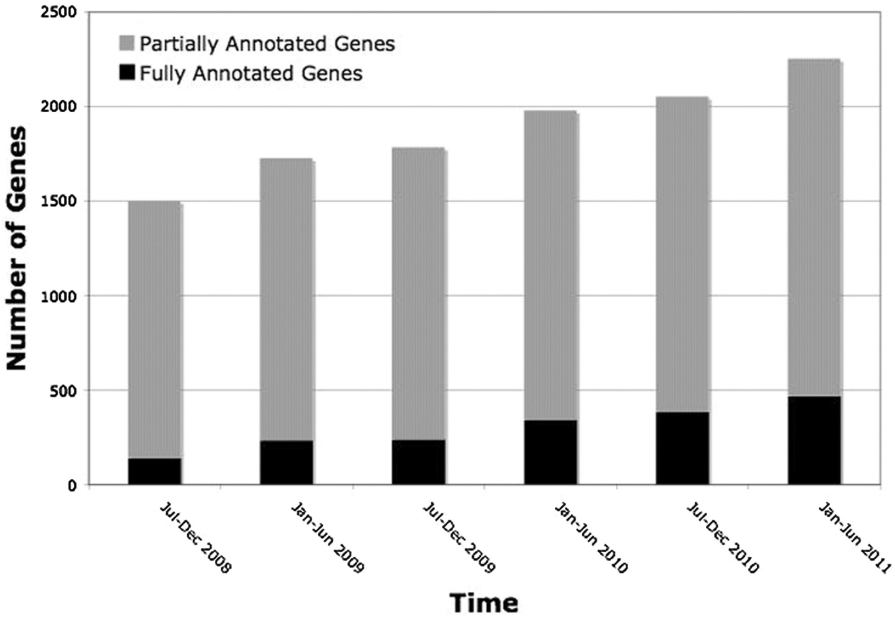

Figure 2. Progress toward annotation of *E. coli* genes. EcoliWiki and EcoCyc are collaborating to improve how the experimental literature is captured in GO annotations (8). Complete annotation of a gene is defined as at least one literature-based annotation to each of the three ontologies: molecular function, cellular component and biological process; unreviewed computational annotations (those with the IEA evidence code) are omitted. The gene association file with current annotations is available at http://www.geneontology.org/GO.downloads.annotations.shtml.

properties and links to information at other sites. Each of these is seeded and updated with information that users can elaborate via community annotation. A section is provided to allow users to place a Jmol structure viewer (http://www.jmol.org/) on the product page. EcoliWiki includes instructions for how to use Jmol to make Proteopedia-style (27) links that launch scripts to customize the structure image.

The *Expression* page contains information about gene expression and regulation. Transcription units are shown with an image of the operon provided by RegulonDB (1). The *Expression* page includes a table for quantitative data about the level, synthesis and degradation of the gene product and/or its mRNA. The *Evolution* page has information about homologs and includes links to sequence comparison and ortholog databases such as BLAST (16), InParanoid (28), CDD (29) and YOGY (30).

Literature

A common concern about wiki-based content is its credibility. An important way of improving the credibility of community-provided content is to provide a robust system for linking assertions to the peer-reviewed primary literature. EcoliWiki uses a modified version of the Cite extension used by Wikipedia. Our modified version recognizes 'ref' tags containing PubMed identifers in the format '<ref name = 'PMID:#'/>'. EcoliWiki uses these tags to generate in-place numbered citation links and a reference list with bibliographic information provided by NCBI's E-Utilities (31). The reference also generates a link to create a page about the cited article in EcoliWiki. This reference page is named for the PMID and is seeded with information from PubMed including the abstract, links to the full-text article (if available) and a table for GO annotations relevant to that publication. Reference pages allow for a community-driven equivalent of the curated papers functionality of other model organism databases.

Other page types

Table 2 lists major page types in EcoliWiki. EcoliWiki aims to provide information about any topic of interest to scientists working with laboratory *E. coli*, its phage, plasmids and mobile elements, either from interest in the *E. coli*'s fundamental biology or simply as a biotechnology tool. Thus, we have pages for cloning vectors, strains, methods and other online resources. EcoliWiki generates pages for any GO term used in an annotation; these can be used for information about how a particular process or complex is used in the organisms covered by EcoliWiki.

One of the most important aspects of wikis is how they allow users to create new pages. EcoliWiki has a system of form-based entry that will create several page types based on internal templates. Users can also create pages that do not necessarily fit into any predefined type.

INTEGRATION WITH INTERNAL AND EXTERNAL TOOLS

One of the early design decisions for EcoliWiki was whether or not to build an independent website or incorporate the desired *E. coli* content into Wikipedia, as has been done by the GeneWiki project (32). Although other factors contributed to our decision to build a stand-alone website, a major benefit is in how it allows us to customize the wiki to handle complex data in structured tables. The

Table 2. Major types of pages in EcoliWiki

Page type	Number of pages
Gene-centric pages	37 362
Strains	419
Plasmids	347
Databases	205
Bioinformatics tools	68
Methods	66
Education & Teaching Tools	22
Literature	23 906
Cis Elements	23
Mobile Genetic Elements	61
Transcription Units	4164
Promoters	31
General Information (biology, complexes, genome sequencing projects, phantom genes, genes not mapped to a genome sequence)	51
User pages	763
Infrastructure (templates, notices, etc.)	10 695
Total (including Talk pages, genes, etc.)	154 887

division of pages into sections of tables and notes makes EcoliWiki more complex than some biological wikis, and it has been criticized for that complexity (7). However, this structure provides us with the ability to retrieve data from the wiki without the need for natural language processing to extract specific types of information from free text. This allows us to integrate EcoliWiki content with internal and external tools in ways that would otherwise be much more difficult (Supplementary Data: Technical and Table S1).

Within the wiki itself, data from tables on other pages is used in the *Quickview* table. The *Quickview* table not only looks up nomenclature from the *Gene* and *Gene Product(s)* pages, but also uses coordinates and sequences to provide data to linked tools for displaying the sequence or viewing alternative reading frames using EMBOSS (33), designing PCR primers with GBrowse (25) and Primer3 (34), and engineering silent restriction sites (35). On the *Gene Product(s)* page, a graphical display of the locations of domains and mutations is generated from information in two different tables on the *Gene* and *Gene Product(s)* pages.

As noted above, EcoliWiki *Gene* pages incorporate thumbnail images of local genome context using our installation of GBrowse (25). Our instance of GBrowse provides genome browsing for all the *E. coli* and phage genomes in EcoliWiki, and for several plasmids and mobile elements. We have added several custom tracks to the *E. coli* MG1655 GBrowse, including transcription units from RegulonDB (1), rRNA operons, cryptic prophage, repetitive elements and others. We have been adding tracks for data from ChIP-chip studies available from public repositories or from authors (36–41). These tracks link to EcoliWiki reference pages for more detailed descriptions of the experiments and data processing.

EcoliWiki has implemented a modified version of Textpresso (42), a full-text literature search engine. Our Textpresso periodically updates its corpus from reference information added to EcoliWiki. In 2011, EcoliWiki deployed a tool that allows users to search fitness data and correlations from a large-scale phenotyping study published by Nichols *et al.* (43). This tool uses EcoliWiki to match gene names and synonyms to genes used in the study.

Structured data allows EcoliWiki to provide web services for the integrated PortEco search of different *E. coli* resources. EcoliWiki web services can also be used to identify pages that have been edited in a specified date range.

DISCUSSION

EcoliWiki and other wiki-based resources show that wiki software can be adapted for many of the purposes of an online model organism database. EcoliWiki adds other freely available tools such as GBrowse and Jmol, and our own open-source development to extend the capabilities of the basic Mediawiki platform.

Both web analytics (Figure 3) and anecdotal feedback from users suggest that EcoliWiki is successful in the sense that it is widely used by our target community, and is viewed as a useful resource. As a community annotation system, however, our goals also include encouraging users to contribute to the content. Any user can view content on the EcoliWiki website; however, users must register in order to create or edit content on the site. Although the requirement for an account presents an unfortunate disincentive for community participation, requiring account creation is a relatively quick and simple method to avoid spam and vandalism. We use a 'vampire model' for user registration, where any registered user can create new users.

Non-staff users have contributed 1513 edits to 485 pages since its rollout in 2007 as of 10 August 2011. While this is encouraging, the majority of the manually curated content in EcoliWiki is still generated by members of the EcoliWiki project. Only a small fraction of the user base has contributed content. EcoliWiki is viewed by several thousand users who are identified as repeat visitors (e.g. 6668 visited at least 5 times between 11 August 2010 and 10 August 2011), but there are only 710 non-staff registered users. Of these, only 170 have edited EcoliWiki, and many of these have only edited it once (Figure 3).

Increased user participation is needed to realize one of the basic ideas of the wiki model: that quality is improved by multiple users reviewing and refining the same content. To further encourage community participation, we display contributor usernames for each page on the sidebar in addition to the standard display of editors in the page history. We also promote editing the wiki in workshops and through email contact with authors of recent papers. Nevertheless, in the absence of other incentives, the low editing participation is consistent with what is seen in other resources built on voluntary collaboration, including Wikipedia (http://stats.wikimedia.org/EN/Sitemap.htm). Wikipedia overcomes this by having such a large user base that even a small fraction of users are sufficient to create and improve millions of pages of content.

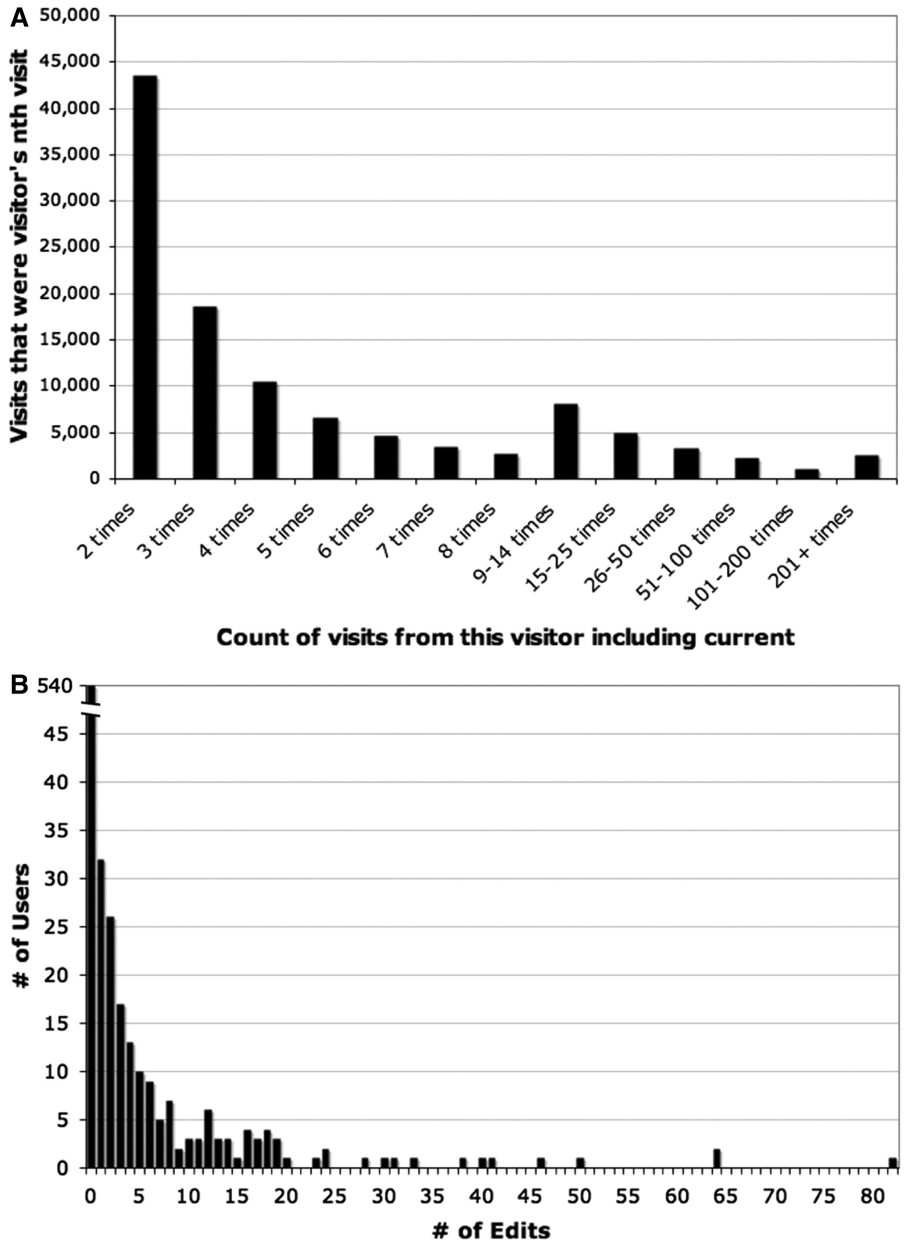

Figure 3. (**A**) Visitor statistics from Google Analytics for 11 August 2010–10 August 2011. The numbers of visits that are the *n*-th visit for a given user indicate users who return to EcoliWiki multiple times. Note that the number of users who visited *n* times is included in the count for users who visited fewer than *n* times. Thus, each bar is for the number of users who visited *n* times or more. (**B**) Distribution of user contributions to EcoliWiki, excluding project employees.

Increasing the participation in editing will require new ways to make editing align with the other incentives that determine how scientists allocate their scarce resources of time and energy. One potential mechanism is to encourage academic scientists to incorporate editing EcoliWiki into teaching. A section of the wiki on Educational Resources can be used to share materials and methods for this purpose. We have also developed wiki extensions to allow instructors to more easily track where their students have edited EcoliWiki.

Future directions for EcoliWiki will also focus on increasing its utility through further integration with other tools via the PortEco project (http://porteco.org). In particular, we are working to improve the connections with our PortEco partners, EcoCyc (3), the PortEco instance of the Stanford Microarray Database (44) and PANTHER (45) in ways that optimize their synergy with wikis. For example, while it would not be appropriate to have community editing of data from a published transcriptome experiment, community curation of the metadata associated with that experiment can be valuable for analyses across experiments from different labs. Similarly, while it might be difficult to devise a system for direct community editing of phylogenetic trees, community

commentary could enrich the evaluation of hypotheses generated from phylogenetic inference.

DATABASE AVAILABILITY

EcoliWiki is freely available via the EcoliWiki website (http://ecoliwiki.net).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Methods.

ACKNOWLEDGEMENTS

We would like to thank Yasha Hartberg and Jerry Tsai for inspiring our initial experiments with wikis for annotation. We thank Lincoln Stein and Mike Cherry for hosting J.C.H. during early development of EcoliWiki and for valuable insights into model organism databases and community annotation. Peter Karp generously helped us use EcoCyc to seed EcoliWiki with content about *E. coli* genes. Barry Wanner and Paul Thomas, as PIs of the EcoliHub and PortEco projects provided the subcontracts to support this work. Dave Clements helped contribute documentation and debugging. Jim Wilson donated the code for the project and Drew Endy graciously shared his laboratory's edited/updated version of the T7 GenBank file. Gunnar von Heijne provided an updated version of TMHMM. We especially want to thank Jim Anderson at National Institutes of Health for deciding that wikis were a promising avenue for data resources, the biological wiki community for generously sharing their experience, information and code, and the users and contributors to EcoliWiki content.

FUNDING

EcoliWiki is funded as a component of PortEco from subcontracts from grant U24 GM077905-01 (2006-2009) and 1U24GM088849-01 (2009-present) from the NIH/NIGMS. Funding for open access charge: NIH 1U24GM088849-01.

Conflict of interest statement. None declared.

REFERENCES

1. Gama-Castro,S., Salgado,H., Peralta-Gil,M., Santos-Zavaleta,A., Muniz-Rascado,L., Solano-Lira,H., Jimenez-Jacinto,V., Weiss,V., Garcia-Sotelo,J.S., Lopez-Fuentes,A. *et al.* (2011) RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res.*, **39**, D98–D105.
2. Glasner,J.D., Liss,P., Plunkett,G. III, Darling,A., Prasad,T., Rusch,M., Byrnes,A., Gilson,M., Biehl,B., Blattner,F.R. *et al.* (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.*, **31**, 147–151.
3. Keseler,I.M., Collado-Vides,J., Santos-Zavaleta,A., Peralta-Gil,M., Gama-Castro,S., Muniz-Rascado,L., Bonavides-Martinez,C., Paley,S., Krummenacker,M., Altman,T. *et al.* (2011) EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res.*, **39**, D583–D590.
4. Misra,R.V., Horler,R.S., Reindl,W., Goryanin,II. and Thomas,G.H. (2005) EchoBASE: an integrated post-genomic database for Escherichia coli. *Nucleic Acids Res.*, **33**, D329–D333.
5. Rudd,K.E. (2000) EcoGene: a genome sequence database for Escherichia coli K-12. *Nucleic Acids Res.*, **28**, 60–64.
6. Medigue,C., Viari,A., Henaut,A. and Danchin,A. (1993) Colibri: a functional data base for the Escherichia coli genome. *Microbiol. Rev.*, **57**, 623–654.
7. Florez,L.A., Roppel,S.F., Schmeisky,A.G., Lammers,C.R. and Stulke,J. (2009) A community-curated consensual annotation that is continuously updated: the Bacillus subtilis centred wiki SubtiWiki. *Database*, **2009**, bap012.
8. Hu,J.C., Karp,P.D., Keseler,I.M., Krummenacker,M. and Siegele,D.A. (2009) What we can learn about Escherichia coli through application of Gene Ontology. *Trends Microbiol.*, **17**, 269–278.
9. Lee,T.L. (2008) Big data: open-source format needed to aid wiki collaboration. *Nature*, **455**, 461.
10. Hu,J.C., Aramayo,R., Bolser,D., Conway,T., Elsik,C.G., Gribskov,M., Kelder,T., Kihara,D., Knight,T.F. Jr, Pico,A.R. *et al.* (2008) The emerging world of wikis. *Science*, **320**, 1289–1290.
11. Hayashi,K., Morooka,N., Yamamoto,Y., Fujita,K., Isono,K., Choi,S., Ohtsubo,E., Baba,T., Wanner,B.L., Mori,H. *et al.* (2006) Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. *Mol. Syst. Biol.*, **2**, 2006 0007.
12. Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F. *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277**, 1453–1462.
13. Durfee,T., Nelson,R., Baldwin,S., Plunkett,G. III, Burland,V., Mau,B., Petrosino,J.F., Qin,X., Muzny,D.M., Ayele,M. *et al.* (2008) The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. *J. Bacteriol.*, **190**, 2597–2606.
14. Ferenci,T., Zhou,Z., Betteridge,T., Ren,Y., Liu,Y., Feng,L., Reeves,P.R. and Wang,L. (2009) Genomic sequencing reveals regulatory mutations and recombinational events in the widely used MC4100 lineage of Escherichia coli K-12. *J. Bacteriol.*, **191**, 4025–4029.
15. Jeong,H., Barbe,V., Lee,C.H., Vallenet,D., Yu,D.S., Choi,S.H., Couloux,A., Lee,S.W., Yoon,S.H., Cattolico,L. *et al.* (2009) Genome sequences of Escherichia coli B strains REL606 and BL21(DE3). *J. Mol. Biol.*, **394**, 644–652.
16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. Demerec,M., Adelberg,E.A., Clark,A.J. and Hartman,P.E. (1966) A proposal for a uniform nomenclature in bacterial genetics. *Genetics*, **54**, 61–76.
18. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.
19. UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
20. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
21. Scheer,M., Grote,A., Chang,A., Schomburg,I., Munaretto,C., Rother,M., Sohngen,C., Stelzer,M., Thiele,J. and Schomburg,D. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res.*, **39**, D670–D676.
22. Kiefer,F., Arnold,K., Kunzli,M., Bordoli,L. and Schwede,T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.*, **37**, D387–D392.
23. Pieper,U., Webb,B.M., Barkan,D.T., Schneidman-Duhovny,D., Schlessinger,A., Braberg,H., Yang,Z., Meng,E.C., Pettersen,E.F., Huang,C.C. *et al.* (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.*, **39**, D465–D474.

24. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
25. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
26. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
27. Hodis,E., Prilusky,J. and Sussman,J.L. (2010) Proteopedia: A collaborative, virtual 3D web-resource for protein and biomolecule structure and function. *Biochem. Mol. Biol. Educ.*, **38**, 341–342.
28. Ostlund,G., Schmitt,T., Forslund,K., Kostler,T., Messina,D.N., Roopra,S., Frings,O. and Sonnhammer,E.L. (2010) InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res.*, **38**, D196–D203.
29. Marchler-Bauer,A., Lu,S., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.
30. Penkett,C.J., Morris,J.A., Wood,V. and Bahler,J. (2006) YOGY: a web-based, integrated database to retrieve protein orthologs and associated Gene Ontology terms. *Nucleic Acids Res.*, **34**, W330–W334.
31. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
32. Huss,J.W. 3rd, Lindenbaum,P., Martone,M., Roberts,D., Pizarro,A., Valafar,F., Hogenesch,J.B. and Su,A.I. (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.*, **38**, D633–D639.
33. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
34. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
35. Reidhaar-Olson,J.F., Bowie,J.U., Breyer,R.M., Hu,J.C., Knight,K.L., Lim,W.A., Mossing,M.C., Parsell,D.A., Shoemaker,K.R. and Sauer,R.T. (1991) Random mutagenesis of protein sequences using oligonucleotide cassettes. *Methods Enzymol.*, **208**, 564–586.
36. Sanchez-Romero,M.A., Busby,S.J., Dyer,N.P., Ott,S., Millard,A.D. and Grainger,D.C. (2010) Dynamic distribution of SeqA protein across the chromosome of Escherichia coli K-12. *mBio*, **1**, e00012–10.
37. Cho,H., McManus,H.R., Dove,S.L. and Bernhardt,T.G. (2011) Nucleoid occlusion factor SlmA is a DNA-activated FtsZ polymerization antagonist. *Proc. Natl Acad. Sci. USA*, **108**, 3773–3778.
38. Cho,B.K., Zengler,K., Qiu,Y., Park,Y.S., Knight,E.M., Barrett,C.L., Gao,Y. and Palsson,B.O. (2009) The transcription unit architecture of the Escherichia coli genome. *Nat. Biotechnol.*, **27**, 1043–1049.
39. Grainger,D.C., Hurd,D., Goldberg,M.D. and Busby,S.J. (2006) Association of nucleoid proteins with coding and non-coding segments of the Escherichia coli genome. *Nucleic Acids Res.*, **34**, 4642–4652.
40. Mooney,R.A., Davis,S.E., Peters,J.M., Rowland,J.L., Ansari,A.Z. and Landick,R. (2009) Regulator trafficking on bacterial transcription units in vivo. *Mol. Cell*, **33**, 97–108.
41. Vora,T., Hottes,A.K. and Tavazoie,S. (2009) Protein occupancy landscape of a bacterial genome. *Mol. Cell*, **35**, 247–253.
42. Muller,H.M., Kenny,E.E. and Sternberg,P.W. (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.*, **2**, e309.
43. Nichols,R.J., Sen,S., Choo,Y.J., Beltrao,P., Zietek,M., Chaba,R., Lee,S., Kazmierczak,K.M., Lee,K.J., Wong,A. *et al.* (2011) Phenotypic landscape of a bacterial cell. *Cell*, **144**, 143–156.
44. Hubble,J., Demeter,J., Jin,H., Mao,M., Nitzberg,M., Reddy,T.B., Wymore,F., Zachariah,Z.K., Sherlock,G. and Ball,C.A. (2009) Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Res.*, **37**, D898–D901.
45. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.